A systematic phylogenetic approach to study the interaction of HIV-1 with coinfections, non-communicable and opportunistic diseases

Katharina Kusejko[1,2], Nadine Bachmann[1,2], Sandra E. Chaudron[1,2], Huyen Nguyen[1,2], Dominique L. Braun[1,2], Benjamin Hampel[1,2], Manuel Battegay[3], Enos Bernasconi[4], Alexandra Calmy[5], Matthias Cavassini[6], Matthias Hoffmann[7], Jürg Böni[2], Sabine Yerly[5], Thomas Klimkait[8], Matthieu Perreau[9], Andri Rauch[10], Huldrych F. Günthard[1,2*], Roger D. Kouyos[1,2*], and the Swiss HIV Cohort Study

1 Division of Infectious Diseases and Hospital Epidemiology, University Hospital Zurich, Zurich, Switzerland **2** Institute of Medical Virology, University of Zurich, Zurich, Switzerland **3** Division of Infectious Diseases and Hospital Epidemiology, University Hospital Basel, University of Basel, Basel, Switzerland **4** Division of Infectious Diseases, Regional Hospital Lugano, Lugano, Switzerland **5** Laboratory of Virology and Division of Infectious Diseases, Geneva University Hospital, University of Geneva, Geneva, Switzerland **6** Division of Infectious Diseases, Lausanne University Hospital, Lausanne, Switzerland **7** Division of Infectious Diseases, Cantonal Hospital St Gallen, St. Gallen, Switzerland **8** Molecular Virology, Department of Biomedicine–Petersplatz, University of Basel, Basel, Switzerland **9** University of Lausanne, Lausanne, Switzerland **10** Department of Infectious Diseases, Bern University Hospital, University of Bern, Bern, Switzerland

*Shared authorship.

Key words: HIV, coinfections, comorbidities, opportunistic infections, phylogenetic analysis;

Short summary:

Systematically analyzing the clustering of clinical endpoints on the HIV-phylogeny at a population level gives insight into the syndemic nature of HIV with other coinfections and non-communicable diseases, as well as virus traits potentially relevant for certain diseases.

Abstract

To systematically test whether coinfections spread along the HIV-1 transmission network and whether similarities of HIV-1 genomes predict AIDS-defining illnesses and comorbidities, we analyzed the distribution of these variables on the HIV-phylogeny of the densely sampled Swiss HIV Cohort Study. By combining different statistical methods, we could detect, quantify and explain the clustering of diseases: Infectious conditions such as hepatitis C, but also Kaposi's sarcoma, clustered significantly, suggesting transmission of these infections along the HIV-1 transmission network. The clustering of patients with neurocognitive complaints, however, could not be completely explained by the clustering of patients with similar demographic risk factors, which suggests a potential impact of viral genetics. In summary, the consistent and robust signal for infectious conditions highlights the strong interaction of HIV-1 and other infections and shows the potential of combining phylogenetic methods to identify disease traits that are likely to be related to virus genetic factors.

Background

The life expectancy of people with HIV (PWH) under successful antiretroviral therapy (ART) is approaching the life expectancy of HIV-1 uninfected individuals (1,2). However, several studies have shown that PWH suffer more often from a broad spectrum of comorbidities compared to the uninfected population (3–5). The increased risk for comorbidities in PWH can possibly be explained by the elevated inflammation due to the HIV-1 infection, direct consequences of the HIV-1 infection for the immune system, the long-term toxicity of ART, as well as social risk factors and lifestyle (6).

Managing comorbidities and coinfections in the growing and ageing population of PWH remains challenging for infectious disease physicians. In Switzerland, PWH are screened regularly and routinely for a broad range of diseases, but factors causing specific diseases in PWH are only incompletely understood. In order to improve the targeting of screening programs and more detailed examinations, such as in-depth neurological screening or bone density measurements, more studies on the distribution of comorbidities in PWH are necessary.

In this project, we systematically analyze all clinically relevant coinfections, non-communicable and opportunistic diseases reported in the Swiss HIV Cohort Study (SHCS). We use a phylogenetic tree to infer the HIV-1 transmission network based on clusters of sequences within the tree. We study the distribution of clinical endpoints across the clusters. A non-random distribution of patients suffering from similar diseases can have several implications: 1) Coinfections could share transmission routes with HIV, 2) Patients could have similar underlying social networks 3) Direct influence of viral genetic factors. By including possible clinical, behavioral and lifestyle factors for each disease, we aim to disentangle the effects of these three causes (see **Figure 1** for the work flow). This approach of using the HIV-phylogeny to understand viral traits was so far mainly applied to study the heritability of set-point viral load and CD4 decline (7–9) and for studying imprinting of neutralization responses against HIV-1 by similar viruses (10). Transmission of coinfections along the HIV-1 transmission network was to date only studied for hepatitis C (HCV) (11,12).

In this work, we expand this type of analysis in a systematic manner to all clinically relevant coinfections, non-communicable and opportunistic diseases collected in the SHCS.

Methods

Swiss HIV Cohort Study

The SHCS, launched in 1988, is a prospective multi-center cohort study enrolling diagnosed HIV-infected adults in Switzerland ([http://www.shcs.ch]) (13). For all participants, demographic information is collected at baseline, laboratory and behavioral information in half-yearly follow-up visits. The SHCS was approved by the ethics committees of the participating institutions and written informed consent was obtained from all participants.

Definitions

In **Table 1**, we define all analyzed coinfections and non-communicable diseases. HIV-related opportunistic diseases were defined, according to the Center for Disease Control, as stage B or stage C infections. See **S3** for more detailed information about all studied clinical endpoints. We included four demographic variables: transmission group, sex, age and ethnicity, three clinical variables: time on ART, CD4 nadir and HCV coinfection and four lifestyle variables: smoking, body mass index (BMI), hypertension and condom use (**S1**). To correct for a potential calendar time bias, all time-dependent variables were taken either at the time of diagnosis of the disease of interest, or at a systematically elaborated reference date for patients not having the disease (**S2.1**). Variable inclusion for the different diseases was performed by a systematic algorithm (**S2.2**).

Study population

The population at risk is defined as the subset of patients for whom the disease of interest was either clearly diagnosed, or clearly not present. Therefore, the size of the study population changes for every condition (see **Table 1**). For opportunistic diseases, the study population is the set of patients who had at least one opportunistic disease. All diseases were studied separately, not including information on the co-occurrence of several diseases in the same patient: A preliminary analysis did not reveal high correlations between any two conditions (**S2.5**).

Phylogenetic tree, clusters and cherries

Phylogenetic tree: A maximum-likelihood phylogenetic tree was built using sequences of 11,915 patients from the genotypic-resistance-test database of the SHCS and non-Swiss background sequences from the Los Alamos database. See **S2.6** for details on the tree construction and earlier SHCS projects using this approach (7,14).

Clusters: All clusters with at least 80% SHCS sequences were extracted from the tree, called Swiss clusters. We concentrated on Swiss clusters having a maximal pairwise cophenetic distance of 0.045. Since distance cut-offs for inclusion of clusters used in HIV-1 phylogenetic studies range between 0.01 and 0.045 (15), we performed extensive sensitivity analyses on this threshold (**S3**).

Cherries: In addition to the analysis of clusters of any size, we concentrated on clusters of size 2, i.e., pairs of SHCS sequences that share a direct common ancestor and are potential transmission pairs, called cherries. Again, only pairs with a maximal cophenetic distance of 0.045 were considered (see **S3** for sensitivity analyses).

In **S2.3**, we report the size and number of clusters and cherries by distance threshold (**S2.3.1**), as well as the distribution of different transmission groups (**S2.3.2**) and subtypes (**S2.3.3**) across clusters and cherries.

Statistical analysis

For the analysis of each disease, only patients in the respective study population were included (see **Table 1**). We used two different methods to understand the distribution of patients suffering from similar diseases on the tree. Method 1: A mixed effects logistic regression model *(16)* was used to analyze whether patients suffering from a particular disease are randomly distributed across the clusters or not, i.e., whether they cluster on the tree. The dependent variable describes whether the disease of interest was present or not, the phylogenetic clusters were included as random effect. With a likelihood ratio test we could test whether including this random effect significantly improved the model fit. Method 2: A parent-offspring regression was applied on the cherries to quantify the odds of having a disease if the other patient in the cherry has the disease as well. See **S2.4** for a detailed explanation of the models used. We performed univariable and multivariable analyses for both methods: Several demographic and clinical factors associated with the disease of interest were included as fixed effects in the mixed effects model as well as the parent-offspring model (7). Risk factors were included according to a systematic algorithm (**S2.2**). See **Figure 1** for a summary of the work flow. All analyses were performed with R (version 3.4.4).

Results

Study population

The phylogenetic tree consists of sequences of 11,915 SHCS patients and 11,390 Los Alamos background sequences. The phylogenetic cluster analysis revealed that 7,195 (60%) patients were in Swiss clusters with cophenetic distance of less than 0.045. Moreover, 5,244 (44%) patients were in a cherry with cophenetic distance less than 0.045. **Table 1** displays the numbers of patients at risk for each studied condition. **Table 2** summarizes

information about the variables included in the multivariable analyses, i.e., the factors potentially associated with different diseases.

In the following, we analyze three types of conditions: Coinfections, non-communicable diseases, and opportunistic diseases. The results of the first method (mixed effects model) are reported by the p values of the likelihood ratio test performed using clusters with distance threshold 0.045. The quantitative results of the second method (parent-offspring regression) are reported by the odds ratios (OR) and 95% confidence intervals (CI) obtained for the analysis of cherries with distance threshold 0.045 (see **S3** for alternative distance thresholds).

Coinfections

The coinfections HCV, hepatitis B (HBV), syphilis, cytomegalovirus (CMV) and latent tuberculosis were analyzed. Of the patients included in the corresponding analyses, 19.9% were HCV coinfected, 32.8% HBV coinfected, 24.0% had at least once syphilis, 85.6% were coinfected with CMV and 9.0% were coinfected with latent tuberculosis (see **Table 1**). All five infections clustered significantly on the phylogenetic tree when applying the mixed effect model ($p < 0.001$ for HCV, HBV, syphilis and CMV; $p = 0.011$ for latent tuberculosis). In the parent-offspring regression, the impact of the neighbor in the cherry being coinfected or not was highest for HCV (OR = 9.5, CI: [7.3, 12.5]), followed by syphilis (OR = 3.0, CI: [2.4, 3.7]), latent tuberculosis (OR = 2.2, CI: [1.3, 3.5]), CMV (OR = 2.0, CI: [1.5, 2.6]) and HBV (OR = 1.4, CI: [1.1, 1.8]). After adjusting for risk factors, all coinfections except latent tuberculosis remained significant in the mixed effects approach. See **Figure 2** for a summary and **S3.1** for detailed information on the selection of confounders and sensitivity analysis.

Non-communicable diseases

Chronic kidney disease (CKD): CKD was diagnosed in 15.1% of the study population. Patients with CKD clustered significantly on the tree ($p < 0.001$), even after adjustment for age, time on ART, CD4 nadir, HCV coinfection and hypertension. The odds of having CDK if the other patient in the cherry has CKD were significantly increased (OR = 1.4, CI: [1.0, 2.0]), however, not after adjusting for risk factors. As a sensitivity analysis, we included tenofovir (TDF) as a potential risk factor, which is known to be associated with lower eGFR (17), but this did not change the results (see **S3.2**).

Cardiovascular diseases: There were 1,778 cardiovascular events of 961 patients in the study population (**S3.2**) which is 6.3% of the study population. The mixed effects model showed that patients with cardiovascular diseases were not randomly distributed on the tree ($p < 0.001$), the clustering however disappeared when correcting for sex and transmission group, age, smoking and hypertension. No increased odds were observed in the parent-offspring regression (OR = 1.3, CI = [0.6,2.4]).

Diabetes mellitus: Patients with diabetes, which constituted 6.1% of the study population, clustered significantly on the phylogenetic tree ($p < 0.001$), again with no significant clustering when correcting for sex and transmission group, age, years on ART, CD4 nadir, BMI and hypertension. We did not find increased odds for having diabetes in the case the other patient in the cherry had diabetes (OR = 1.5, CI = [0.7, 2.9]).

Osteoporosis: A total of 14.4% of the study population had osteoporosis. Neither the mixed effects model nor the parent-offspring regression revealed significant clustering. In the selection algorithm for the multivariable analysis, only low BMI was selected as a risk factor, possibly due to the small sample size (see **Table** 1).

Neurocognitive complaints: Neurocognitive questions about frequent memory loss, concentration problems and slowing down in reasoning revealed that 9.8% of the study

population had neurocognitive complaints. Patients with complaints clustered significantly on the phylogenetic tree ($p < 0.001$) with increased odds of having complaints if the other patient in the cherry had complaints (OR: 2.0, CI: [1.2, 3.4]). After adjusting for sex and transmission group, age, years on ART, smoking and BMI, we still observed significant clustering ($p = 0.004$) and increased odds (OR: 1.9, CI: [1.1, 3.3]).

Psychiatric events: Psychiatric events were recorded for 38.5% of the study population. Patients with psychiatric events clustered significantly on the tree ($p < 0.001$), however not after adjustment for sex and transmission group, age, years on ART, HCV coinfection, smoking, BMI, hypertension and condom use. The parent-offspring regression revealed significant results both in the univariable model (OR: 1.6, CI: [1.3, 1.9]) as well as multivariable model (OR: 1.4, CI: [1.1, 1.7]).

Non-HIV associated neoplasms: Patients with neoplasms, which constituted 5.8% of the study population, clustered significantly on the tree ($p < 0.001$), even after adjusting for age, years on ART, CD4 nadir and smoking. The odds of having a neoplasm if the other patient in the cherry had a neoplasm were increased (OR: 2.0, CI: [1.1, 3.4]), however not after adjustment for risk factors (OR: 1.5, CI: [0.8, 2.7]).

More information on all non-communicable diseases, including details on the variable selection and sensitivity analysis on the distance threshold, can be found in **S3.2**. **Figure 2** summarizes the results on non-communicable diseases.

Opportunistic diseases

There were 4,528 patients in the SHCS phylogeny with at least one opportunistic disease. Due to a small number of patients in clusters and even fewer in cherries (see **Table 1**), we could only use the mixed effects method for analyzing phylogenetic clustering of patients with frequently observed opportunistic diseases. In addition, we corrected for the potential risk factors age, ethnicity, ART and CD4 nadir only one by one. See **Table 3** for a

summary of the analysis. Patients who suffered from candida stomatitis ($p = 0.025$), weight loss ($p = 0.001$), HIV-related encephalopathy ($p = 0.004$), Kaposi's sarcoma ($p < 0.001$), bacterial pneumonia ($p < 0.001$) and cervical dysplasia ($p < 0.001$) were not randomly distributed across the clusters. After correcting for age, ethnicity, ART and CD4 nadir, respectively, the clustering remained significant for some of these conditions. In particular, patients who suffered from candida stomatitis, weight loss, Kaposi's sarcoma, bacterial pneumonia, and cervical dysplasia still clustered significantly on the tree. For HIV-related encephalopathy, the clustering remained significant after correcting for age and ethnicity, respectively, but not for the intake of ART and CD4 nadir.

Discussion

In this study, we used the HIV-phylogeny of the SHCS to study patterns of the occurrence of comorbidities, coinfections, and HIV-related illnesses. Proximity of patients on the phylogeny can have several implications: First, these patients are close in the HIV-1 transmission network and hence similar transmission routes are likely. Second, patients who are close on the phylogeny might in addition share a social network and are therefore more likely to have a similar lifestyle. Third, proximity of patients translates to proximity of the viral genome.

The coinfections HCV, HBV, syphilis, CMV and latent tuberculosis all clustered significantly on the tree. The odds of being HCV-coinfected was, e.g., 9.5-times higher if the other patient in the cherry was HCV-coinfected. A slightly higher odds ratio was found by Kouyos et al (11), who used a similar parent-offspring approach in the SHCS, however, without restrictions on cophenetic distance. This suggests a syndemic nature of these five infections and HIV-1. Interestingly, clustering of patients coinfected with latent tuberculosis could be explained by demographic confounding, meaning that patients originating from high prevalence countries for tuberculosis also more likely shared similar HIV strains. This is in line with Fenner et al (18), who showed that HIV infection disrupts the sympatric host-

pathogen relationship (adaptation of a pathogen to a host population resulting in co-evolution) in human tuberculosis in Switzerland, meaning that in the HIV population in Switzerland, allopatric tuberculosis strains are mainly found. Contrariwise, Koch et al (19) found evidence for an impact of HIV on the evolution of tuberculosis in South Africa. This difference highlights that our results on coinfections are setting specific: Tuberculosis is a rare disease in Switzerland and the main risk group are PWH coming from countries with a high tuberculosis prevalence (20) or intravenous drug users (IDU) (21). Accordingly, tuberculosis transmission in Switzerland is rare (in contrast to other coinfections including syphilis and hepatitis C), which is in line with the weak clustering observed in our study. Tuberculosis prevalence is however high in South Africa, with frequent community and household transmission and the HIV-negative population being possibly disproportionally responsible for onward transmission (22).

Several opportunistic diseases clustered significantly on the phylogeny: The clustering of patients with Kaposi's sarcoma may indicate shared transmission routes of HIV-1 and human herpesvirus-8, the pathogen causing Kaposi's sarcoma (23). Although patients with HIV-related encephalopathy clustered not any more when correcting for CD4 nadir and intake of ART, clustering was significant when correcting for age and ethnicity, respectively. CD4 nadir and intake of ART can, from a clinical perspective, not completely explain the clustering of patients suffering from HIV-related encephalopathy among AIDS-patients, as a low CD4 nadir and no treatment are risk factors for all AIDS-related diseases. Thus, this result is a tentative indicator of neuropathogenic traits of some HIV-1 strains. This supports previous suggestions that more neuropathogenic HIV-1 strains may exist (24). Our results could be viewed as indicators of additional pathogenesis traits, but more evidence is needed to strengthen this hypothesis.

For most non-communicable diseases analyzed, patients were likewise not distributed randomly on the phylogeny. In most cases, however, the clustering was not significant in the multivariable analysis, which means that most of the variables we corrected

for are likewise not distributed randomly on the tree. This is expected for variables such as the transmission group, as HIV-1 is most often transmitted among risk groups (25). Clustering of behavioral factors such as smoking or condom use highlights that the HIV-phylogeny does not only represent the HIV-1 transmission network but also the underlying social network. Correcting for age, BMI, hypertension and smoking could explain the clustering of, e.g., patients with cardiovascular events or diabetes. The fact that clustering vanished for most analyzed non-communicable diseases in the multivariable model reflects the rich and detailed data on potential confounders available in the SHCS. It suggests that this data can capture the clustering of diseases caused by clustering of socio-demographic and behavioral factors. In addition, adjustment for confounding suggests that these non-communicable diseases are not influenced much by viral genetic traits.

Clustering of patients with psychiatric problems and neurocognitive complaints disappeared in the multivariable model, but the odds of having psychiatric problems and neurocognitive complaints if the other patient in the cherry had these problems remained significantly increased (see **Figure 2**). We used patients' self-reported neurocognitive complaints as a proxy for neurocognitive problems. Simioni et al (26) showed that these complaints correlate well with symptomatic forms of HIV-associated neurocognitive disorders (HAND), but asymptomatic forms might not be detected. The results obtained in our analysis might therefore become stronger when looking at in-depth neurocognitive screening for both asymptomatic and symptomatic forms of HAND. Similar to the clustering of patients with HIV-related encephalopathy, the clustering of neurocognitive complaints could hence be seen as a tentative indicator that certain HIV-1 strains are more prone to damage the brain.

Our results could be used for developing more targeted screening programs, e.g., in-depth neurological screening or testing for syphilis. For conditions which were not randomly distributed on the phylogeny, clinicians could decide based on the viral sequence whether a patient should undergo additional screening: if the patients' sequence is in a cluster with a high prevalence of the disease of interest, screening is advisable. Including viral sequences

in selecting patients for in-depth screening programs could thus expand the current approach of choosing patients based on demographic, clinical or behavioral aspects.

Our study has several strengths and limitations. One strength is that HIV-1 sequence data was available for 11,915 patients. The SHCS provides in addition to HIV-related clinical and laboratory information in-depth information about coinfections and non-communicable diseases. Thus, we were able to perform a population-based study systematically investigating a variety of diseases and could choose from a selection of different epidemiological and clinical risk factors. One drawback is, however, that some of the considered phenotypes, e.g., cardiovascular diseases or non-HIV associated neoplasms, summarize a heterogeneous spectrum of diseases with different etiologies and different risk factors correlated with these conditions. In addition, some of the studied conditions and covariables rely on self-reported information given by the patient and others were evaluated only on a biased subset of patients, e.g., osteoporosis. In all analyses, we used two different approaches to describe the distribution of diseases on the HIV-phylogeny, in addition to different cophenetic distance thresholds. In several cases, the different methods and thresholds led to slightly different results which makes interpretation difficult (**S2.4**). However, we performed extensive sensitivity analysis to understand the impact of the different approaches and thresholds (**S3**). Based on this sensitivity analysis, we conclude that the results presented here are robust and clustering of comorbidities is rather underestimated for liberal distance thresholds. This result of obtaining stronger clustering for more conservative distance thresholds was also described earlier (27). Moreover, our results proved to be robust when restricting the analysis to patients infected with HIV subtype B (**S4**). Nevertheless, we want to emphasize that out results should be viewed as tentative indicators of shared transmission routes or viral genetic impact with more evidence and further research needed to prove these hypotheses. Our analysis could provide candidates for more targeted search of the effect of individuals' genes or mutations. Moreover, we would like to highlight that our results, especially the quantitative results on shared transmission routes of HIV with other coinfections, are specific for the Swiss and other western European settings,

where white men who have sex with men (MSM) comprise the main risk group for onward transmission of HIV.

In conclusion, this work for the first time presents a systematic analysis interrogating the HIV-phylogeny at a population level for the syndemic nature of coinfections and non-communicable diseases, respectively for virus traits potentially relevant for certain diseases. The large variety of conditions tested is a strength of this project, implies however that no universal explanation or interpretation of the clustering can be given. There is evidence for three different reasons for clustering: shared transmission routes of pathogens, similar social networks of patients close in the phylogeny and direct viral genetic impact. Overall, our strategy, together with adjustment for numerous known confounding factors, demonstrates the potential for a new type of analysis, extending conventional epidemiological analyses.

Additional information (Foot note page)

Acknowledgements

We thank the patients who participate in the Swiss HIV Cohort Study; the physicians and study nurses, for excellent patient care; the resistance laboratories, for high-quality genotyping drug resistance testing; SmartGene (Zug, Switzerland), for technical support; Alexandra Scherrer, Susanne Wild, Anna Traytel from the SHCS data center for data management, Danièle Perraudin and Marianne Amstutz for administration. The members of the Swiss HIV Cohort Study include the following:

Anagnostopoulos A, Battegay M, Bernasconi E, Böni J, Braun DL, Bucher HC, Calmy A, Cavassini M, Ciuffi A, Dollenmaier G, Egger M, Elzi L, Fehr J, Fellay J, Furrer H (Chairman of the Clinical and Laboratory Committee), Fux CA, Günthard HF (President of the SHCS), Haerry D (deputy of "Positive Council"), Hasse B, Hirsch HH, Hoffmann M, Hösli I, Huber M, Kahlert C, Kaiser L, Keiser O, Klimkait T, Kouyos RD, Kovari H, Ledergerber B, Martinetti G, Martinez de Tejada B, Marzolini C, Metzner KJ, Müller N, Nicca D, Paioni P, Pantaleo G, Perreau M, Rauch A (Chairman of the Scientific Board), Rudin C (Chairman of the Mother & Child Substudy), Scherrer AU (Head of Data Centre), Schmid P, Speck R, Stöckle M, Tarr P, Trkola A, Vernazza P, Wandeler G, Weber R, Yerly S.

Ethics statement

The SHCS was approved by the ethics committees of the participating institutions (Kantonale Ethikkommission Bern, Ethikkommission des Kantons St. Gallen, Comité Départemental d'Éthique des Spécialités Médicales et de Médicine Communataire et de Premier Recours, Kantonale Ethikkommission Zürich, Repubblica et Cantone Ticino–Comitato Ethico Cantonale, Commission Cantonale d'Éthique de la Recherche sur l'Être Humain, Ethikkommission beider Basel for the SHCS and Kantonale Ethikkommission Zürich for the ZPHI), and written informed consent was obtained from all participants.

Funding

This work was supported by the Swiss National Science Foundation (Grant # BSSGI0_155851). HFG was supported by SNF grant 179571. Furthermore, this study has been financed within the framework of the Swiss HIV Cohort Study, supported by the Swiss National Science Foundation (grant #148522), by the SHCS research foundation by the Yvonne Jacob Foundation (to HFG), by the clinical research priority program of the University of Zurich “Viral infectious diseases, ZPHI” (to HFG). HFG has received an unrestricted research Grant from Gilead to the SHCS Research Foundation.

Conflicts of interests

HFG has received unrestricted research grants from Gilead Sciences and Roche; fees for data and safety monitoring board membership from Merck; consulting/advisory board membership fees from Gilead Sciences, Sandoz and Mepha; and travel reimbursement from Gilead. MB has received research or educational grants by Abb Vie AG, Gilead Sciences Switzerland Sàrl, Janssen-Cilag AG, MSD Merck Sharp & Dohme AG and ViiV Healthcare GmbH. EB has received fees for his institution for participation to advisory board from MSD, Gilead Sciences, ViiV Healthcare, Abbvie and Janssen. MC has received research and travel grants for his institution from ViiV and Gilead. AC has received unrestricted educational and research grants from MSD, Gilead and ViiV. AR reports support to his institution for advisory boards and/or travel grants from Janssen-Cilag, MSD, Gilead Sciences, Abbvie, and Bristol-Myers Squibb, and an unrestricted research grant from Gilead Sciences. All remuneration went to his home institution and not to AR personally, and all remuneration was provided outside the submitted work.

Meeting(s) where the information has been presented

The content of this work was presented at CROI (Conference on Retroviruses and Opportunistic Infections) in March 2018 in Boston (Poster number 169: Phylogenetic clusters of HIV-1 reveal potential viral genetic impact on comorbidities).

Author contact information

Corresponding author:

Katharina Kusejko, PhD

Division of Infectious Diseases and Hospital Epidemiology

University Hospital Zürich, Rämistrasse 100, CH-8091 Zürich

+41 43 253 0188

katharina.kusejko@usz.ch

Alternate corresponding author:

Roger Kouyos, PhD

Division of Infectious Diseases and Hospital Epidemiology

University Hospital Zürich, Rämistrasse 100, CH-8091 Zürich

+41 43 255 3610

roger.kouyos@usz.ch

References

1. Trickey A, May MT, Vehreschild J-J, Obel N, Gill MJ, Crane HM, et al. Survival of HIV-positive patients starting antiretroviral therapy between 1996 and 2013: a collaborative analysis of cohort studies. Lancet HIV. 2017 Aug 1;4(8):e349–56.

2. Gueler A, Moser A, Calmy A, Günthard HF, Bernasconi E, Furrer H, et al. Life expectancy in HIV-positive persons in Switzerland: matched comparison with general population. AIDS Lond Engl. 2017 Jan 28;31(3):427–436.

3. Butt AA, McGinnis K, Rodriguez-Barradas MC, Crystal S, Simberkoff M, Goetz MB, et al. HIV infection and the risk of diabetes mellitus. AIDS Lond Engl. 2009 Jun 19;23(10):1227–1234.

4. Triant VA, Brown TT, Lee H, Grinspoon SK. Fracture prevalence among human immunodeficiency virus (HIV)-infected versus non-HIV-infected patients in a large U.S. healthcare system. J Clin Endocrinol Metab. 2008 Sep;93(9):3499–504.

5. Triant VA, Lee H, Hadigan C, Grinspoon SK. Increased acute myocardial infarction rates and cardiovascular risk factors among patients with human immunodeficiency virus disease. J Clin Endocrinol Metab. 2007 Jul;92(7):2506–12.

6. Hasse B, Ledergerber B, Furrer H, Battegay M, Hirschel B, Cavassini M, et al. Morbidity and Aging in HIV-Infected Persons: The Swiss HIV Cohort Study. Clin Infect Dis. 2011 Dec 1;53(11):1130–9.

7. Bachmann N, Turk T, Kadelka C, Marzel A, Shilaih M, Böni J, et al. Parent-offspring regression to estimate the heritability of an HIV-1 trait in a realistic setup. Retrovirology. 2017 May 23;14(1):33.

8. Alizon S, von Wyl V, Stadler T, Kouyos RD, Yerly S, Hirschel B, et al. Phylogenetic approach reveals that virus genotype largely determines HIV set-point viral load. PLoS Pathog. 2010 Sep 30;6(9):e1001123.

9. Blanquart F, Wymant C, Cornelissen M, Gall A, Bakker M, Bezemer D, et al. Viral genetic variation accounts for a third of variability in HIV-1 set-point viral load in Europe. PLoS Biol. 2017 Jun;15(6):e2001855.

10. Kouyos RD, Rusert P, Kadelka C, Huber M, Marzel A, Ebner H, et al. Tracing HIV-1 strains that imprint broadly neutralizing antibody responses. Nature. 2018;561(7723):406–10.

11. Kouyos RD, Rauch A, Böni J, Yerly S, Shah C, Aubert V, et al. Clustering of HCV coinfections on HIV phylogeny indicates domestic and sexual transmission of HCV. Int J Epidemiol. 2014 Jun;43(3):887–896.

12. Vanhommerig JW, Bezemer D, Molenkamp R, Van Sighem AI, Smit C, Arends JE, et al. Limited overlap between phylogenetic HIV and hepatitis C virus clusters illustrates the dynamic sexual network structure of Dutch HIV-infected MSM. AIDS Lond Engl. 2017 Sep 24;31(15):2147–2158.

13. Schoeni-Affolter F, Ledergerber B, Rickenbach M, Rudin C, Günthard HF, Telenti A, et al. Cohort Profile: The Swiss HIV Cohort Study. Int J Epidemiol. 2010 Oct 1;39(5):1179–1189.

14. Turk T, Bachmann N, Kadelka C, Böni J, Yerly S, Aubert V, et al. Assessing the danger of self-sustained HIV epidemics in heterosexuals by population based phylogenetic cluster analysis. eLife. 2017 Sep 12;6.

15. Ragonnet-Cronin M, Hodcroft E, Hué S, Fearnhill E, Delpech V, Brown AJL, et al. Automated analysis of phylogenetic clusters. BMC Bioinformatics. 2013 Nov 6;14:317.

16. Fitting Linear Mixed-Effects Models Using lme4 | Bates | Journal of Statistical Software. [cited 2018 Aug 13]; Available from: https://www.jstatsoft.org/article/view/v067i01

17. Quesada PR, Esteban LL, García JR, Sánchez RV, García TM, Alonso-Vega GG, et al. Incidence and risk factors for tenofovir-associated renal toxicity in HIV-infected patients. Int J Clin Pharm. 2015 Oct;37(5):865–72.

18. Fenner L, Egger M, Bodmer T, Furrer H, Ballif M, Battegay M, et al. HIV infection disrupts the sympatric host-pathogen relationship in human tuberculosis. PLoS Genet. 2013;9(3):e1003318.

19. Koch AS, Brites D, Stucki D, Evans JC, Seldon R, Heekes A, et al. The Influence of HIV on the Evolution of Mycobacterium tuberculosis. Mol Biol Evol. 2017 01;34(7):1654–68.

20. Sudre P, Hirschel B, Toscani L, Ledergerber B, Rieder HL. Risk factors for tuberculosis among HIV-infected patients in Switzerland. Swiss HIV Cohort Study. Eur Respir J. 1996 Feb;9(2):279–83.

21. Stucki D, Ballif M, Bodmer T, Coscolla M, Maurer A-M, Droz S, et al. Tracking a tuberculosis outbreak over 21 years: strain-specific single-nucleotide polymorphism typing combined with targeted whole-genome sequencing. J Infect Dis. 2015 Apr 15;211(8):1306–16.

22. Middelkoop K, Mathema B, Myer L, Shashkina E, Whitelaw A, Kaplan G, et al. Transmission of tuberculosis in a South African community with a high prevalence of HIV infection. J Infect Dis. 2015 Jan 1;211(1):53–61.

23. Gramolelli S, Schulz TF. The role of Kaposi sarcoma-associated herpesvirus in the pathogenesis of Kaposi sarcoma. J Pathol. 2015 Jan;235(2):368–80.

24. Pillai SK, Pond SLK, Liu Y, Good BM, Strain MC, Ellis RJ, et al. Genetic attributes of cerebrospinal fluid-derived HIV-1 env. Brain J Neurol. 2006 Jul;129(Pt 7):1872–83.

25. Kouyos RD, von Wyl V, Yerly S, Böni J, Taffé P, Shah C, et al. Molecular epidemiology reveals long-term changes in HIV type 1 subtype B transmission in Switzerland. J Infect Dis. 2010 May 15;201(10):1488–1497.

26. Simioni S, Cavassini M, Annoni J-M, Rimbault Abraham A, Bourquin I, Schiffer V, et al. Cognitive dysfunction in HIV patients despite long-standing suppression of viremia. AIDS. 2010 Jun;24(9):1243.

27. Kusejko K, Kadelka C, Marzel A, Battegay M, Bernasconi E, Calmy A, et al. Inferring the age difference in HIV transmission pairs by applying phylogenetic methods on the HIV transmission network of the Swiss HIV Cohort Study. Virus Evol [Internet]. 2018 Jul 1 [cited 2018 Sep 26];4(2). Available from: https://academic.oup.com/ve/article/4/2/vey024/5101427

28. CKD stages - The Renal Association [Internet]. [cited 2018 Aug 8]. Available from: https://renal.org/information-resources/the-uk-eckd-guide/ckd-stages/

29. Levey AS, Stevens LA, Schmid CH, Zhang YL, Castro AF, Feldman HI, et al. A new equation to estimate glomerular filtration rate. Ann Intern Med. 2009 May 5;150(9):604–12.

30. Genuth S, Alberti KGMM, Bennett P, Buse J, Defronzo R, Kahn R, et al. Follow-up report on the diagnosis of diabetes mellitus. Diabetes Care. 2003 Nov;26(11):3160–7.

31. Anagnostopoulos A, Ledergerber B, Jaccard R, Shaw SA, Stoeckle M, Bernasconi E, et al. Frequency of and Risk Factors for Depression among Participants in the Swiss HIV Cohort Study (SHCS). PLoS ONE [Internet]. 2015 Oct 22 [cited 2017 Jul 30];10(10). Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4619594/

Supporting Information

S1 Potential risk factors

S2 Methods

S3 Diseases

S4 Sensitivity Analysis: Subtype B restriction

Figures and Tables

Fig 1*: Flow chart to visualize the work flow. Step 1: For each disease, only patients with the respective information about the disease were included. Step 2: The reference dates for the control population, i.e., the patients not suffering from the disease of interest, were selected. Step 3: Variables which are potentially correlated with the disease of interest were selected. Step 4: Mixed effect models and parent-offspring regression was used to describe the distribution of patients on the tree suffering from the same disease. Step 5: Clustering was interpreted and studied in more detail in a sensitivity analysis.*

***Table 1**: Synopsis of all study populations for all analyzed coinfections, non-communicable diseases and opportunistic infections including the respective definitions of the comorbidities. First and second column: The group of diseases and singular diseases studied, respectively, Third column: Definitions of the singular diseases studied, Fourth column: The number of SHCS patients included in the phylogenetic tree who underwent testing for the respective diseases of interest. Fifth column: The subset of SHCS patients who are in any cluster (distance cut-off 0.045) (i.e. patients who underwent testing for the respective comorbidities and are in a phylogenetic cluster). Sixth column: The subset of SHCS patients who are in a cherry (distance cut-off 0.045) (i.e. patients who underwent testing for the respective comorbidities and are in a cherry).*

Groups of diseases studied	**Singular diseases studied**	**Definition**	**Size of the study populations**		
			All SHCS patients	**All patients in a cluster**	**All patients in a cherry**
Coinfections	Hepatitis C	A positive HCV antibody test or positive HCV RNA.	9,710	5,417	3,584
	Hepatitis B	A positive anti-HBc antibody test.	8,325	4,462	2,770
	Syphilis	A positive venereal disease research laboratory or rapid plasma reagin test confirmed by a treponema specific test.	9,635	5,411	3,556
	Cytomegalovirus infection	A positive result of the CMV immunoglobulin G test.	11,506	6,851	4,852

	Latent tuberculosis	Latent tuberculosis was defined as a positive tuberculin skin reactivity or a positive result of interferon-based screening test.	9,301	5,223	3,258
Non-communicable diseases	Chronic kidney disease	An estimated glomerular filtration rate (eGFR) of less than 60 ml/min/1.73m^2 for two consecutive measurements at least three months apart (28). The eGFR was calculated with the CKD-EPI formula by Leyer et al (29), which considers differences in sex and ethnicity (see Supplementary Material, Section 3.2).	10,248	5,887	3,946
	Cardio-vascular disease	History of coronary angioplasty, myocardial infarction, procedures on other arteries, cerebral infarction and other clinically diagnosed cardiovascular diseases (see Supplementary Material, Section 3.2).	10,933	6,386	4,426
	Diabetes mellitus	A fasting plasma glucose molarity of at least 7 mmol/l on two consecutive occasions or at least 11.1 mmol/l for not fasting (30), where fasting was defined as no caloric intake at least eight hours before the blood sample was taken, as well as for patients who took antidiabetic agents.	10,715	6,235	4,292
	Osteoporosis	Bone mineral density was measured with dual-energy X-ray absorptiometry (DXA) for a subset of patients. Patients with a test value	2,295	760	290

		of less than minus 2.5 times the standard deviation of a comparable healthy person of either of the measured sites, namely hip, lumbar and neck, were diagnosed with osteoporosis.			
	Neurocognitive complaints	Neurocognitive complaints were assessed using three questions about memory loss, concentration problems and slowing down in reasoning in the follow-up questionnaire completed semiannually by all patients (26). Only problems in all three categories were considered as neurocognitive problems.	8,046	4,272	2,638
	Psychiatric problems	Hospitalization for psychiatric reasons, suicide, self-reported depression that was either diagnosed by a psychiatrist or other physician (31) or self-reported depression with intake of antidepressants (Supplementary Material, Section 3.2.6).	9,080	5,008	3,236
	Non-HIV associated neoplasms	All non-HIV associated neoplasms recorded in the SHCS.	11,687	7,015	5,046

Opportunistic infections		All stage B and C infections according to the Center for Disease Control.	4,538	2,070	1,010

Table 2: *Summary of the most important host demographic and laboratory information: The second column shows the characteristics of all patients in the phylogenetic tree, the third column of all patients who are in a cluster (distance cut-off 0.045) and the fourth column of all patients who are in a cherry (distance cut-off 0.045). For time dependent parameters, e.g., years on antiretroviral treatment (ART), we present here information at the last follow-up of the patients; MSM: men who have sex with men, HET: heterosexuals, IDU: intravenous drug users, IQR: interquartile range, BMI: body mass index;*

	All patients in the phylogenetic tree	**All patients in a cluster**	**All patients in a cherry**
Sex and transmission group			
MSM	4736 (39.7%)	2977 (41.4%)	2108 (40.2%)
Male HET	1952 (16.4%)	1095 (15.2%)	844 (16.1%)
Female HET	2292 (19.2%)	1074 (14.9%)	885 (16.9%)
Male IDU	1579 (13.3%)	1184 (16.5%)	794 (15.1%)
Female IDU	850 (7.1%)	620 (8.6%)	426 (8.1%)
Male Other	305 (2.6%)	169 (2.3%)	124 (2.4%)
Female Other	200 (1.7%)	75 (1%)	62 (1.2%)
Year of birth (median, IQR)	1965 (1959 - 1972)	1965 (1959 - 1971)	1965 (1959 - 1971)
Ethnicity			
white	9199 (77.2%)	6052 (84.1%)	4325 (82.5%)

		All patients in the phylogenetic tree	All patients in a cluster	All patients in a cherry
non-white		2716 (22.8%)	1143 (15.9%)	919 (17.5%)
ART naive		811 (6.8%)	508 (7.1%)	363 (6.9%)
Years on ART				
	>5	8308 (69.7%)	5011 (69.6%)	3657 (69.7%)
	3-5	1083 (9.1%)	660 (9.2%)	480 (9.2%)
	1-3	1186 (10%)	700 (9.7%)	515 (9.8%)
	<1	504 (4.2%)	305 (4.2%)	220 (4.2%)
	no	834 (7%)	519 (7.2%)	372 (7.1%)
CD4 nadir (median, IQR)		181 (71 - 294)	188 (80 - 298)	188 (78 - 299)
Hepatitis C coinfection		2999 (25.2%)	2161 (30%)	1476 (28.1%)
Smoking		6746 (56.6%)	4429 (61.6%)	3177 (60.6%)
BMI (median, IQR)		26 (23 - 28)	26 (23 - 28)	26 (23 - 28)
Hypertension		4671 (39.2%)	2806 (39%)	2071 (39.5%)
Condomless anal intercourse		3158 (26.5%)	2088 (29%)	1464 (27.9%)

Fig 2: *Summary of the analysis for coinfections and non-communicable diseases: "Method 1" displays p values of the likelihood ratio test used in the mixed effects model approach (Method 1) on all clusters. We show the result of the univariable model (UV) for the distance (d) cut-off 0.045, as well as the results of the multivariable (MV) approach for distance cut-offs 0.045, 0.035, 0.025, 0.015. "Method 2" displays the odds ratios obtained in the parent-offspring regression approach on all cherries, again UV with distance 0.045 and MV for distance cut-offs 0.045, 0.035, 0.025, 0.015.*

Table 3: *Summary of the results of the cluster analysis for opportunistic infections: The second column ("Frequency") denotes the frequency of the respective disease among all opportunistic diseases recorded in the SHCS; The third and fourth column display the p values of the likelihood ratio test applied to the mixed effects model approach (Method 1) on all clusters, for the univariable analysis (UV) and for the multivariable analysis (MV) after adjusting for the factors potentially associated with the clustering.*

Opportunistic disease	Frequency	p value (UV)	p value (MV)
Candida stomatitis	44.6%	0.023	Age: 0.017, Ethnicity: 0.033, ART naive: 0.02, CD4 nadir: 0.039
Oral hairy leukoplakia	24.3%	0.046	-
Candidiasis esophageal	19.0%	no sign.	-
HIV-related thrombocytopenia	17.1%	no sign.	-
Herpes zoster multidermatomal or relapse	16.0%	no sign.	-
Pneumocystis pneumonia	14.6%	no sign.	-
Weight loss	9.4%	0.001	Age: < 0.001, Ethnicity: 0.002,

Opportunistic disease	Frequency	p value (UV)	p value (MV)
			ART naive: 0.004, CD4 nadir: 0.005
HIV-related encephalopathy	9.2%	0.004	Age: < 0.001, Ethnicity: 0.005, ART naive: 1, CD4 nadir: 0.432
Kaposi's sarcoma	8.6%	< 0.001	Ethnicity: < 0.001, ART naive: < 0.001,
Toxoplasmosis of the brain	7.7%	no sign.	-
Recurrent bacterial pneumonia	7.6%	< 0.001	Age: < 0.001, Ethnicity: < 0.001, ART naive: < 0.001, CD4 nadir: < 0.001
Aids wasting syndrome	6.6%	no conv.	-
Mucocutan. Herpes simplex ulceration	6.3%	no conv.	-
CMV retinitis	5.3%	< 0.001	-
Non-Hodgkin's lymphoma	4.9%	no sign.	-
Disseminated MAC disease	4.7%	no sign.	-

Opportunistic disease	Frequency	p value (UV)	p value (MV)
Cervical dysplasia	4.7%	0.001	Age: 0.036, Ethnicity: 0.002, ART naive: 0.009, CD4 nadir: 0.002
M. tuberculosis (pulmonary and extrapulmonary)	4.6%	no conv.	-
CMV not liver spleen or lymph nodes	3.7%	no sign.	-
Cryptosporidiosis	2.8%	no sign.	-
Progressive multifocal leukoencephalopathy	2.7%	no sign.	-

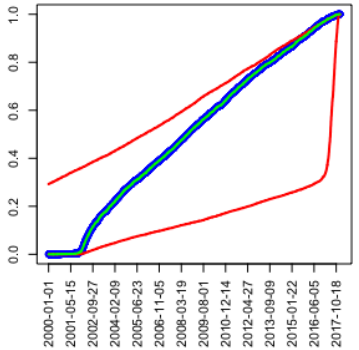
Figure 1

Step 1: Definition of the study population

Only patients with information about the disease of interest, e.g., available laboratory measurements, are included.
Patients with ambiguous results are excluded.

Step 2: Choice of reference dates

Cases: The date the disease was diagnosed for the first time.
Controls: A date between the first and the last time point the disease was not present is chosen such that the distribution over time is the same for cases and controls.
→ Supplements S2.1



Step 3: Variable selection

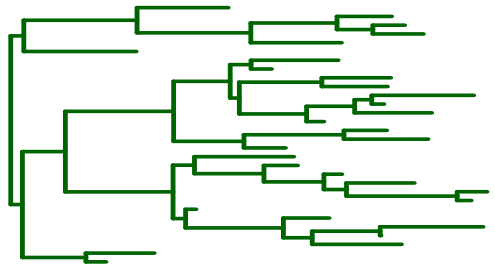
Covariables (*Sex, transmission group, age, ethnicity, ART, CD4, HCV, Smoking, BMI, Hypertension, Condom use*) are selected based on data availability and significance in the univariable and multivariable fixed model (without cluster information).
→ Supplements S2.2

.

Step 4: Mixed effects model and parent-offspring regression

$$\text{disease} \sim \text{intercept} + \beta_1 \text{variable}_1 + ... + \beta_N \text{variable}_N + (1|\text{clusterNR})$$

$$\text{disease}_{\text{offspring}} \sim \text{intercept} + \beta_0 \text{disease}_{\text{parent}} + \beta_1 \text{variable}_1 + ... + \beta_N \text{variable}_N$$



→ Supplements S2.4

Step 5: Interpretation and sensitivity analysis

Clustering of diseases is interpreted based on their infectious or non-infectious nature. A sensitivity analysis on the cophenetic distance threshold is performed for all diseases.
→ Supplements S3

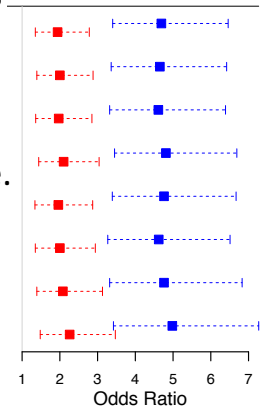


Figure 2

Summary of the analysis for coinfections and non-communicable diseases

	Method 1: Mixed effects model					Method 2: Parent-offspring regression
	UV: d = 0.045	MV: d = 0.045	MV: d = 0.035	MV: d = 0.025	MV: d = 0.015	UV: d = 0.045 MV: d = 0.045 MV: d = 0.035 MV: d = 0.025 MV: d = 0.015
Coinfections						
Hepatitis C	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	
Hepatitis B	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	
Syphilis	< 0.001	0.127	0.036	0.043	0.123	
Cytomegalovirus	< 0.001	< 0.001	0.034	0.051	0.178	
Latent tuberculosis	0.011	0.588	0.257	0.674	0.024	
Non–communicable diseases						
Chronic kidney disease	< 0.001	0.006	< 0.001	< 0.001	< 0.001	
Cardio–vascular diseases	< 0.001	0.859	0.252	0.107	< 0.001	
Diabetes mellitus	< 0.001	0.401	1	1	0.376	
Osteoporosis	0.169	0.188	0.253	0.257	0.407	
Neurocognitive complaints	< 0.001	0.004	0.073	0.678	< 0.001	
Psychiatric problems	< 0.001	0.116	0.369	0.493	0.064	
Non–HIV associated neoplasms	< 0.001	0.019	0.014	< 0.001	< 0.001